Comparison of Latent Dirichlet Modeling and Factor Analysis for Topic Extraction: A Lesson of History

Normand Péladeau
Provalis Research
peladeau@provalisresearch.com

Elnaz Davoodi
Concordia University
e_davoo@encs.concordia.ca

Abstract

Topic modeling is often perceived as a relatively new development in information retrieval sciences, and new methods such as Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation have generated a lot of research. However, attempts to extract topics from unstructured text using Factor Analysis techniques can be found as early as the 1960s. This paper compares the perceived coherence of topics extracted on three different datasets using Factor Analysis and Latent Dirichlet Allocation. To perform such a comparison a new extrinsic evaluation method is proposed. Results suggest that Factor Analysis can produce topics perceived by human coders as more coherent than Latent Dirichlet Allocation and warrant a revisit of a topic extraction method developed more than fifty-five years ago, yet forgotten.

1. Introduction

The vast majority of information in any business or organization is unstructured data, typically in text format such as reports, forms, emails, memos, log entries, transcripts, etc. The rapid growth of social media and the digitalization of archived documents further increases the volume of text data available. However, most of the time, this rich source of information remains untapped because of the tremendous effort it takes to sift through and dig out information.

Various exploratory text mining techniques may be used to automatically extract information and find patterns and relationships in large amount of text data. Topic Modeling (TM) is a fast-growing area of research which has recently gained a lot of attention in the text mining community (e.g. [4, 12, 33, 37]). Topic modeling not only can be useful for the end user by discovering recurrent patterns of co-occurring words (i.e. topics), but also can be beneficial for other Natural Language Processing (NLP) applications including sentiment analysis [32], information retrieval [35], text summarization [13], etc. While TM is described by many researchers as a recent technique with references to the development of Latent Dirichlet Allocation (LDA) in 2003 [4], others will associate its true origin to applications of Latent Semantic Analysis (LSA) for the extraction of topics in the late 1990s [18, 30].

However, this idea of automatically extracting topics from unstructured text collection is not new. For example, research in information retrieval as early as 1963 used Factor Analysis (FA) on text documents to extract topics and automatically classify documents [5, 6]. Whilst this work received a lot of attention as an unsupervised approach to document classification, though rarely has it been cited as an example of topic identification. At about the same time, FA was used to automatically generate topics stored in the form of content analysis dictionaries [15, 21]. The research led to the development of a computer program called WORDS which was used in psychology to analyze psychotherapeutic interviews [15, 21, 16] and changes in productivity of manic patients [17]. This software was also used to study the humanities and linguistics [25, 23]. The initial efforts in TM were all based on FA, an unsupervised method for discovering latent variables. The same TM technique using FA performed on a Principal Component Analysis (PCA) with varimax rotation is implemented today in WordStat content analysis software[1].

Very few contemporary research articles focusing on TM mention those early efforts. Further, we could not find any systematic attempt to compare new techniques with topics extracted using FA. Without such a comparison, it is hard to know whether more recent approaches to TM represent a real improvement over the original work done fifty-five years ago.

This paper will present the results of such a comparison. The evaluation of the topic models is still an active area of research and suffers from a lack of widely accepted evaluation methods. In this paper, we propose a novel method of conducting a human

[1] https://provalisresearch.com/wordstat

URI: http://hdl.handle.net/10125/49965
ISBN: 978-0-9981331-1-9

evaluation. We first explain two of the most widely used methods for TM (i.e. LSA and LDA) and the original TM method (i.e. FA). The experimental efforts toward a manual evaluation of topic models are presented, followed by the presentation, analysis and discussion of the results.

2. Latent Semantic Analysis

Latent Semantic Analysis [10] was initially introduced in the Information Retrieval (IR) domain to capture the similarity between documents. LSA is used for dimensionality reduction to represent a document using a vector of latent semantic concepts instead of a vector of words. The dimensionality reduction in LSA is obtained by decomposing a large word-document matrix using Singular Value Decomposition (SVD). As a result, the large term-document matrix is composed of a term-concept matrix, a matrix of singular values and a concept-document matrix. In the context of TM, each concept which is an underlying hidden variable, can be considered as a topic. Probabilistic Latent Semantic Analysis (pLSA) [18] is a variation of LSA. In this model, instead of reducing dimensionality of the word-document matrix using SVD, it uses a probabilistic perspective for discovering underlying hidden variables (i.e. topics). This model is based on two main assumptions: (1) there is a distribution of a fixed number of topics for each document (Formula 1), and (2) there is a distribution over fixed size of vocabulary for each topic (Formula 2). Considering V as a fixed size of vocabulary and T as a fixed number of topics, $\theta_{(t,d)}$ represents the probability of topic t occurs in document d and $\phi_{(w,t)}$ represents the probability of term w is generated by topic t, we can formulate the two above-mentioned assumptions as follows:

$$p_{(t|d)} = \theta_{(t,d)} \quad s.t. \quad \sum_{w \in T} \theta_{(t,d)} = 1 \tag{1}$$

$$p_{(w|t)} = \phi_{(w,t)} \quad s.t. \quad \sum_{w \in V} \phi_{(w,t)} = 1 \tag{2}$$

Finally, the probability of a topic for a given document and a given word are generated using tue Bayes rule. The Expectation-Maximization (EM) algorithm is then used to estimate the parameters of this model.

3. Latent Dirichlet Allocation

Latent Dirichlet Allocation [4] is a probabilistic approach to TM which aims to improve pLSA. In pLSA, there is a probability distribution of topics over each document. In other words, each document can be seen as a list of numbers, each denoting the probability of a topic for the document. However, there is no generative probabilistic model for these probabilities. As noted by Blei et al. [4], this could lead to two main problems: (1) the difficulty of assigning these probabilities to documents outside of the training set, and (2) the number of model parameters growing linearly with the size of the corpus. Thus, LDA can be seen as improved pLSA by introducing a Dirichlet prior on document-topic distributions. LDA has been used extensively for TM (e.g. [11, 26, 3, 36, 8]) and various implementations can be found in text mining tools (e.g. tm package in R, LDA-c, Mallet, Gensim).

4. Factor Analysis

In his article, [10] noted that LSA is a variation of FA and called it two-mode factor analysis. FA was initially aimed to reduce the dimensionality of data to discover the latent content from the data [5, 22]. In FA, each word w_i in the vocabulary V containing all words in a corpus, $w_i \in V, \forall i \in \{1, \dots, n\}$, can be represented as a linear function of $m(< n)$ topics (aka *common factors)*, $t_j \in T, \forall j \in \{1, \dots, m\}$. More specifically, Formula 3 shows the representation of each word using common factors (i.e. topics).

$$\begin{aligned} w_1 &= \lambda_{11}t_1 + \lambda_{12}t_2 + \cdots + \lambda_{1m} + t_m + e_1 \\ w_2 &= \lambda_{21}t_1 + \lambda_{22}t_2 + \cdots + \lambda_{2m} + t_m + e_2 \\ &\vdots \\ w_n &= \lambda_{n1}t_1 + \lambda_{n2}t_2 + \cdots + \lambda_{nm} + t_m + e_n \end{aligned} \tag{3}$$

In Formula 3, $\lambda_{ij}, \forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, m\}$ are called factor loading which show the strength of the relationship o**f** each word to each topic. Also, e_i is the error term associated with the representation of each word. The dimensionality reduction in FA is based on the idea that each word is the representation of a linear combination of underlying hidden variables (i.e. topics). Principle Component Analysis [19], is a similar dimensionality reduction to FA with some important difference. The first notable difference is that PCA does not generate a model of underlying principle components similar to the one that FA generates for underlying common factors (see Formula 3). Another notable difference is the flexibility of each approach to the change in the number of dimensions. Both PCA and FA take the number of new dimensions as a hyperparameter. The change in the number of dimensions from m_1 to m_2 for PCA does not affect the m_1 principle components already computed; however, in FA, the model of underlying common factors should be

built again with the new number of dimensions. The last, but not the least difference is related to the representation of the data points. In FA, an error term is always considered, while in PCA principal components are exact linear transformations of the data without considering residual error [23]. Similar to LSA, FA is an effective way for discovering underlying latent semantic concepts; however in both methods, the resulting lower dimensional concepts are difficult to interpret [33].

5. Topic Modeling Evaluation

Intrinsic evaluation of topic models ideally needs a manually annotated corpus with the topics; however, such annotations are very expensive to produce and the gold standard topics reflect the subjectivity in the annotators' topic comprehension. Automatic scoring of topics has been developed to quantify the quality of topic models by measuring the coherence of words in each topic [28, 7, 29, 1, 31]. However, these metrics still need to be confirmed based on users' preference. In addition to the intrinsic evaluation of topic models, an *extrinsic* evaluation serves as a confirmation to quantify the quality of topic models according to human judgment. [9] used two intrusion methodologies to evaluate the topic words and topics. More specifically, they used word intruders and topic intruders to evaluate the quality of topic words and topics respectively. [29] also used human judgment to score the coherence level of each topic in a 3-point scale. Human judgment was then used to evaluate the automatic coherence scoring method by correlating human judgment and different automatic coherence measures.

We propose a novel approach to evaluate topic models. The intuition behind our proposed topic model evaluation is based on the observation that different TM techniques usually generate many topics that are similar enough to be considered highly related, if not equivalent. The proposal is to pair topics generated by different techniques and sharing some similarities and ask human judges in a forced-choice situation to choose the one they perceive as the most coherent and then ask them to score their choice in terms of coherence. All remaining topics that could not be paired are then evaluated independently for their coherence using the same scale.

5.1. Dataset

Three datasets have been used for assessing topic models (see Table 1). The first dataset we used is the TREC AP corpus [14]. This corpus was previously used for TM [4] and is publicly available[2]. The second dataset consists of 1,795 abstracts presented at the Hawaii International Conference on System Sciences (HICSS) between 2014 and 2016. The abstracts are publicly available from the conference website. Finally, the last dataset consists of reviews of twelve hotels in Las Vegas obtained by scraping the Expedia website. This dataset includes 31,898 reviews posted between August 2005 and July 2013 and is available on request from the first author. Table 1 provides basic descriptive statistics about all three datasets.

5.2. Methodology

To conduct the evaluation of the users' preferences between the topics generated by the two approaches, we generated topics using the Mallet implementation for LDA and WordStat text analysis software for FA. For the TREC AP corpus and the hotel review dataset 100 topics were extracted, while 50 topics were obtained from the HICSS abstract dataset. The same custom stop word lists were used in both models, and the original source documents were analyzed as is, without stemming nor lemmatization. In the case of Mallet, the alpha hyperparameter was set to 0.20 and we chose to extract topics after 1000 iterations. The first ten words of each topic were extracted without consideration of their probability, a decision that may negatively affects the coherence of the extracted topic.

Table 1. The statistics of the datasets

	AP corpus	HICSS abstracts	Hotel reviews
# of documents	2,250	1,795	31,898
# of tokens	892,593	236,827	1,758,545
# of types	38,261	11,257	20,114
Type/Token ratio	4.3%	4.8%	1.1%

For WordStat, the analysis was restricted to all words occurring 10 times or more. While the recommended minimum loading value for topic extraction using FA is 0.30 according to [20] or 0.20 as used by [5] this latter criterion resulted in many topics containing fewer than 10 words. The minimum loading criterion was thus reduced to 0.01, allowing for the extraction of 10 words for each topic for all three datasets. This represents a clear violation of the recommended use of FA for TM and it likely negatively affected the coherence of the extracted topics. Table 2 allows one to assess the potential impact of this decision. It clearly shows that the hotel review dataset is likely the most challenging one since only twelve out of a hundred extracted topics had at least ten words reaching a loading of 0.20. The

[2] http://www.cs.princeton.edu/ blei/lda-c

Table 2. WordStat topic modeling results

	AP corpus	HICSS abstracts	Hotel reviews
10 word topics for loading >= 0.20	72 of 100	40 of 50	12 of 100
Average words/topics (loading >=20)	8.4	9.2	6.0
Lowest loading	0.121	0.130	0.027

average number of words per topic was only 6.0. Table 2 also reports the lowest loading needed to obtain the required number of words for all topics in this dataset was 0.027. This is well below the recommended use for FA. The two other datasets appear to be less affected by small loading of words to topics.

Despite such a clear impediment of the quality of extracted topics for the hotel review dataset, we nevertheless decided to include it in the experiment. There is, however, no reason to believe that such a potential lack of topic coherence that likely affected the topics extracted from the hotel review dataset using FA would not likewise affect the results in LDA. In fact, we can assume that a careful examination of word probabilities in Mallet outputs would also have revealed similar issue with this dataset.

Once all words for all topics were obtained from both techniques, topics were paired using Formula 4 described below. The topic matching procedure aims to find correspondence between topics generated by two topic models. Since the order of words in a topic is an indicator of its relevance, our matching formula considers not only the common words between the two topics, but also the position of those words in the list.

$$M_{ab} = \sum_{c \in C} \frac{2 * (n-1) - (POS_{ca} + POS_{cb})}{n * (n-1)} \quad (4)$$

Considering two lists, *a* and *b*, containing *n* topic words, which share a set of common words, called *C*, their matching score M_{ab} is calculated using Formula 4. In this formula, POS_{ca} denotes the position of a common word *c* in list *a* and similarly POS_{cb} denotes the position of a common word *c* in list *b*. The obtained score varies between 0 and 1.

To make it clearer, we show an example of two topics, each contains 10 words, as follows:

A) *data analytics big mining techniques visual events analysis paper sources*

B) *analytics big data techniques visual visualization business making opportunities processing*

The set *C = {big, visual, techniques, data, analytics}* contains common words between the two topics. In this example, n = 10, which is the number of topic words in each topic. POS_{ca} denotes the position of each of the common words in topic A and POS_{cb} denotes the position of the common words in topic B. For example, for the common word "big", its position in topic A is 3 and in topic B is 2. The matching score of these two topics is 0.67.

For the pairing and topic evaluation task, we considered the first 10 words of each topic model. For each dataset, we proceeded to two successive pairings. First, topics obtained from Mallet were used as the reference set, so that each topic was paired with the WordStat topic with the highest similarity score. In some situations, a single WordStat topic could be associated with more than one Mallet topic. In such cases, only the pair with the highest similarity score was retained. We then performed the same operation but using WordStat topics as the reference set. Such a double pairing was needed since paring was not always symmetrical. The topic pairs which appeared in both pairings and reached a minimum criterion of 0.3 were then selected and categorized as comparable topics. All other topics that did not reach this criterion were put aside and categorized as non-comparable topics. Further analyses on which topics were successfully paired and which ones were not, suggest that paired topics are the easiest and most reliable ones to extract. They were the topics obtaining the highest eigenvalues in the FA, as well as those that were the most stable across multiple LDA extractions.

The topic evaluation task for paired topics consists of a forced-choice situation where the evaluator is being asked to choose which topic is the most coherent. To facilitate this evaluation, each topic pair was split into three sets of words: (1) the anchor words consisting of all words common to both topics (2) the list of words unique to Mallet and (3) and those unique to WordStat. For example, the following two lists of topic words:

1) *security information paper cyber policy attacks attack policies threats secure* (Mallet)

2) *malicious attack attacks security threat cyber threats protection privacy detection* (WordStat)

generated this list of common words:

attack attacks cyber security threats

and two lists of unique words:

A) *information paper policies policy secure*

B) *detection malicious privacy protection threat*

The two lists of unique words were randomly presented, and all three sets of words were sorted in alphabetical order to prevent any potential bias toward one approach or the other. The participants had to choose which list of unique words was the most coherent with the list of common words. They could choose either A, B or both if they considered both sets to be equally consistent (or inconsistent). The users were then asked to quantify their confidence in their choice on an ordinal scale.

In a second evaluation task, the participants were asked to assess, on a 4-point rating scale, the coherence of all remaining topics that could not be paired. While words in paired topics were sorted alphabetically, each set of words in this second task were presented in the original order they were extracted by each tool (so theoretically in descending order of topic relevance). Topics from both tools were randomly presented. Participants were then asked to evaluate the coherence level of each topic and quantify their confidence level.

The users were provided instructions for each task and were not allowed to communicate with each other. However, due to the lack of familiarity of some words or topics by some users, additional information was provided on demand to clarify the meaning and/or definition of technical terms.

The number of participants in this evaluation task differs from one experiment to another. In total, 5 subjects evaluated the topics extracted from AP corpus and 11 subjects did the same evaluation for the HICSS abstracts. For the hotel review dataset 4 participants took part in the evaluation of paired topics, and 6 participants assessed the unpaired ones.

6 Results and Analysis

In this section, we present the results of both evaluation tasks for all three datasets. However, an initial look at the extracted topics reveals a significant qualitative difference in the topics produced by the two approaches which needs to be considered. When we look at the top 10 words for all topics, the LDA models tend to include a smaller variety of words. For example, on the hotel review dataset, topics were built using 503 words (out of a maximum of 1,000, i.e. 10 words $*$ 100 topics), with many words occurring in multiple topics such as "hotel" in 34 topics, "room" in 30 topics and "stay" in 22 topics. By comparison, topics extracted by FA returned 826 different words. High frequency words like "room" and "stay" were not part of those, and the most frequent word ("show") was associated with 5 topics. Most words were used only once or twice. Similar results were observed on the two other datasets. For the AP News dataset, the LDA topic model was generated using 576 words in comparison to 889 words for FA. For the HICSS abstract dataset, 344 words were used by LDA to generate 50 topics in comparison to 434 words for the FA topic model.

6.1. Comparable Topic Evaluation Task

Three scores have been used to compare the preference of participants for either one of the TM techniques: (1) the number of votes favoring each solution across all participants and all topics, (2) the number of topics obtaining a majority of votes from participants, and (3) the number of users expressing preference for one TM technique over the other based on their voting patterns across all topics. The decision to present of all three statistics prevents us from drawing conclusions based on either a minority of participants or a small number of topics clearly favoring one technique over the other. Ideally a model should be considered superior if it gets higher scores on all three measures.

Table 3 Statistics of manual evaluation of comparable topic evaluation task

		AP corpus		HICSS abstracts		Hotel reviews	
		WordStat	Mallet	WordStat	Mallet	WordStat	Mallet
All confidence levels	# of votes	142	70	91	60	84	90
	# of topics	35	12	16	10	31	35
	# of users	5	0	11	0	1	3
High confidence levels	# of votes	56	21	49	27	23	23
	# of topics	27	12	15	9	18	16
	# of users	5	0	9	1	1	3

We then filtered the results to focus solely on answers for which the user expressed a confidence level higher than the median score on the confidence scale. Table 3 shows the evaluation results for all paired topics. Results computed on the AP corpus and HICSS abstracts clearly show that topics produced using FA are considered by participants to be more coherent than those obtained using LDA. This is true on all three measures. For example, for the AP corpus, FA obtained more than twice as many votes as LDA (142 vs 70). For 35 of the 47 paired topics a majority of participants considered the topic obtained using FA to be more coherent in contrast to only 12 for the LDA. Although the difference is less drastic for the HICSS corpus, FA still obtained 52% more votes (91 vs 60) and surpassed LDA by 60% in the number of topics considered more coherent by most participants LDA (16 vs 10). This preference cannot be attributed to individual differences since all participants in both experiments considered topics extracted through FA to be more coherent. While the gap between these two techniques is reduced slightly when filtering in only responses where the participant expressed a high level of confidence, the superior perceived coherence of topics using FA remains important and consistent across all three measures.

The hotel review dataset shows, however, results that are much closer, slightly favoring of LDA over FA as expressed by the total number of votes (90 vs 84) and the total number of topics (35 vs 31). Such a lead vanishes when one filters out responses for which the user expressed a low confidence. In both situations, three out of four participants expressed a preference for topics extracted using LDA.

6.2 Non-Comparable Topic Evaluation Task

The non-comparable topic evaluation task aims to evaluate the coherence of topics generated by each TM approach and that could not be paired using the algorithm presented in Section 5.2. To evaluate the coherence of the topics, we conducted an extrinsic evaluation. In this experiment, the topics generated by each model were combined and shuffled first, then the participants were asked to assess the coherence of a topic on a 4-point rating scale. In addition, the participants were asked to score the level of confidence in each of their evaluations. Table 4 shows the average coherence score for topics generated by WordStat (FA) and Mallet (LDA) from all three datasets. The upper part of the table shows the average coherence score for all items, while the bottom half reports this same statistic but only for responses for which users reported a confidence level higher than the median value on the confidence scale.

As seen, when looking at the entire set of responses (all confidence levels), both topic models generate almost equally coherent topics, with a slightly higher score for the LDA on the AP corpus and the HICSS abstracts, while FA generated topics that were judged slightly more coherent than those obtained through LDA for the hotel review dataset. None of the differences were found to be statistically or substantially significant. Considering only responses for which the participants expressed a high level of confidence further reduced those differences.

7. Discussion

By systematically comparing the coherence of topics extracted using WordStat on one side and Mallet on the other side, results clearly suggest that FA has the capability of generating topics that are perceived as more coherent than those obtained through LDA using Mallet. This is despite the common presence of generic words in the latter approach that, in our opinion, may likely reduce the probability of finding words that are incoherent.

FA seems to offer additional benefits over LDA. First, it is well known that the probabilistic nature LDA makes the topic solution subject to variation, which from a user point-of-view, may be perplexing. Generating multiple topic models in LDA will result in different solutions unless an initial random seed value is set. FA, on the other hand, always produces the exact same solution as long as the same options are used. Another possible advantage of FA may be the extraction of topics that are more independent of one another, and potentially provide a more comprehensive description of the text collection. The Varimax rotation is responsible for this since such an orthogonal rotation tends to remove items associated with too many topics and selects items loading strongly on only a few factors (or topics) instead, creating factors that are more independent of each other. On the other hand, topics generated using LDA often contain a smaller variety of words, some of those, especially high frequency ones being associated with numerous topics. As we mentioned before, this presence of generic words like "hotel" and "room" in multiple topics will likely positively affect the perceived coherence of those by human evaluators. However, it also raises more fundamental questions about what should be the desired qualities of topic models. Should the current interest by topic modeling researchers on coherence measured at the topic level be done without considering the specificity of the topics in the entire model? Does the reduced vocabulary that characterizes LDA affects the ability of its topic models to provide a comprehensive

Table 4. Average coherence score for non-comparable topics by levels of confidence

		AP corpus	HICSS abstracts	Hotel reviews
All confidence levels	WordStat	3.26	2.71	3.16
	Mallet	3.33	2.83	3.13
High confidence levels	WordStat	3.32	3.00	3.16
	Mallet	3.36	3.06	3.15

description of the corpus being analyzed? Or contrarily, does the imposition of topic independence or high-specificity through orthogonal rotation in FA create contrived distinctions with no tie to reality? Potential limitations of the present study could be identified and should be further discussed. First, one could very well argue that the topics obtained through LDA were not optimal and that different hyperparameters could have generated more coherent topics. However, the literature on this issue gives very little advice on how such optimization can be achieved. While some have proposed to optimize LDA such that it increases either internal or external coherence measures [28], our attempts to optimize topic modeling this way suggest that such an approach tends to favor topics lacking independence as expressed by even an larger number of high-frequency words overlapping multiple topics. In other words, it appears that topic coherence may well be inversely related to topic specificity.

One may also raise the possibility that more recent algorithms would likely generate more coherent topics. Yet, in light of the obtained results we can argue that such superiority should be established not solely in comparison to more recent algorithms such as pLSA or LDA, but to FA topic modeling as well.

It is crucial to remember that this topic modeling experiment focuses entirely on the descriptive value of those techniques for analyzing unstructured text corpus. For this reason, this study relies exclusively on human judgment of topic coherence. Therefore, it says nothing about the value for FA on other related applications such as document indexing, automatic document classification or information retrieval tasks. It also does not take into consideration important issues such as computational complexity, processing time or processing capability, which are crucial elements when dealing with huge datasets.

We believe that despite the limitations, the results of our experiment clearly plead in favor of the consideration of factor analysis as a legitimate topic modeling technique, especially when used for a descriptive purpose. Further comparative studies involving both FA and LDA as well as more recent topic modeling techniques should be undertaken to identify conditions under which one technique performs better than the others.

The fact that we could not find any contemporary study on topic modeling comparing the performance of techniques such as pLSA or LDA to topic models extracted using FA also raises some legitimate questions about the reason why such a technique is broadly ignored today. Neglect of previous work is not entirely unknown. Already in 1974, Ikers [20] identified at least four instances where researchers reinvented the same technique of FA on word-word association matrices, with no awareness of the others' work. He stressed how paradoxical it was considering that, while the earliest work on this method was done in the area of automatic information retrieval, the technique was consistently rediscovered due to a lack of facilities for automatic index and classification. In his conclusion, he states:

> *"massive amounts of time would have been saved given information of the then current state of affairs in the automatic classification area in the early sixties.[..] One hopes this paper will serve to reduce the information gap. Ikers, p.97.*

We can only make this conclusion ours and reaffirm the importance of re-examining the possible contribution factor analysis could make to the area of topic modeling, a technique that seems to have been overlooked or forgotten (once again) by current researchers.

Acknowledgements

Partial funding of Elnaz Davoodi work was provided by the Mitacs (Mathematics of Information Technology and Complex Systems) program (IT08208). Many thanks also to Ashkan Ebadi for his advices and useful comments while writing this paper and to all the survey participants who took the time from their busy schedules to participate in the study.

References

[1] Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th

International Conference on Computational Semantics (IWCS 2013)–Long Papers. pp. 13–22 (2013)

[2] Berry, M.W., Dumais, S.T., OBrien, G.W.: Using linear algebra for intelligent information retrieval. SIAM review 37(4), 573–595 (1995)

[3] Blei, D.M.: Probabilistic topic models. Communications of the ACM 55(4), 77–84 (2012)

[4] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)

[5] Borko, H., Bernick, M.: Automatic document classification. Journal of the ACM (JACM) 10(2), 151–162 (1963)

[6] Borko, H., Bernick, M.: Automatic document classification part ii. Additional experiments. Journal of the ACM (JACM) 11(2), 138–151 (1964)

[7] Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of the Biennial GSCL Conference. vol. 156 (2009)

[8] Boyd-Graber, J., Blei, D.M.: Multilingual topic models for unaligned text. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. pp. 75–82 (2009)

[9] Chang, J., Boyd-Graber, J.L., Gerrish, S., Wang, C., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Nips. vol. 31, pp. 1–9 (2009)

[10] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science 41(6), 391 (1990)

[11] Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences 101(suppl 1), 5228–5235 (2004)

[12] Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. Psychological review 114(2), 211 (2007)

[13] Haghighi, A., Vanderwende, L.: Exploring content models for multi-document summarization. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 362–370. Association for Computational Linguistics (2009)

[14] Harman, D.: Overview of the second text retrieval conference (trec-2). In: Proceedings of the workshop on Human Language Technology. pp. 351–357 (1994)

[15] Harway, N.I., Iker, H.P.: Computer analysis of content in psychotherapy. Psychological Reports 14(3), 720–722 (1964)

[16] Harway, N.I., Iker, H.P.: Content analysis and psychotherapy. Psychotherapy: Theory, Research & Practice 6(2), 97 (1969)

[17] Harway, N.I., et al.: Some aspects of language style of manic patients. In: Proceedings of the Annual Convention of the American Psychological Association (1973)

[18] Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 50–57 (1999)

[19] Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of educational psychology 24(6), 417 (1933)

[20] Iker, H.P.: An historical note on the use of word-frequency contiguities in content analysis. Computer and the Humanities 8, 93-98 (1974)

[21] Iker, H.P., Harway, N.I.: A computer approach towards the analysis of content. Systems Research and Behavioral Science 10(2), 173–182 (1965)

[22] Iker, H.P., Harway, N.I.: A computer systems approach toward the recognition and analysis of content. The analysis of communication content: Developments in scientific theories and computer techniques pp. 381–405 (1969)

[23] Jandt, F.E.: Sources for computer utilization in interpersonal communication instruction and research. Communication Quarterly 20(2), 25–31 (1972)

[24] Jolliffe, I.: Principal component analysis. Wiley Online Library (2002)

[25] Jonas, T.J.: The WORDS system: a computer-assisted content analysis of Chaim Perelman's "New rhetoric". Ph.D. thesis, Bowling Green State University (1971)

[26] Mcauliffe, J.D., Blei, D.M.: Supervised topic models. In: Advances in neural information processing systems. pp. 121–128 (2008)

[27] Miles, J., Selvin, H.C.: A factor analysis of the vocabulary of poetry in the seventeenth century. The Computer and Literary Style pp. 116–127 (1966)

[28] Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 262–272 (2011)

[29] Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108 (2010)

[30] Papadimitriou, C., Raghavan, P., Tamaki, H., Vempala, S.: Latent Semantic Indexing: A probabilistic analysis. Proceedings of ACM PODS (1998)

[31] Roder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 399–408 (2015)

[32] Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th international conference on World Wide Web. pp. 111–120. ACM (2008)

[33] Wallach, H.M.: Structured topic models for language. Ph.D. thesis, University of Cambridge (2008)

[34] Wang, W., Barnaghi, P.M., Bargiela, A.: Probabilistic topic models for learning terminological ontologies. IEEE Transactions on Knowledge and Data Engineering 22(7), 1028–1040 (2010)

[35] Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 178–185. ACM (2006)

[36] Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 261–270 (2010)

[37] Zeng, J., Zhang, S., Wu, C., Ji, X.: Modeling topic propagation over the internet. Mathematical and Computer Modeling of Dynamical Systems 15(1), 83–93 (2009)

.